For reference:

Zhan, P. (2020). A Markov estimation strategy for longitudinal learning diagnosis: Providing timely diagnostic feedback. *Educational and Psychological Measurement*, Advanced Online Publication, http://doi.org/10.1177/0013164420912318

A Markov Estimation Strategy for Longitudinal Learning Diagnosis: Providing Timely Diagnostic Feedback[*]

Peida Zhan[**]

(College of Teacher Education, Zhejiang Normal University, Jinhua, China)

Abstract

Timely diagnostic feedback is helpful for students and teachers, enabling them to adjust their learning and teaching plans according to a current diagnosis. Motivated by the practical concern that the simultaneity estimation strategy currently adopted by longitudinal learning diagnosis models does not provide timely diagnostic feedback, this study proposes a new Markov estimation strategy, which follows the Markov property. A simulation study was conducted to explore and compare the performance of four estimation strategies: the simultaneity, the Markov, the anchor-item, and the separated estimation strategies. The results show that their performance was highly consistent, and they presented in the following relative order: simultaneity > Markov > anchor-item ≥ separated. Overall, although accuracy in parameter estimation is sacrificed slightly with the proposed strategy, it can provide timely diagnostic feedback to practitioners, which is in line with the concept of "assessment for learning" and the needs of formative assessment.

Keywords: learning diagnosis, Markov property, timely feedback, longitudinal cognitive diagnosis

---

[*] This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No. 19YJC190025), the National Natural Science Foundation of China (Grant No. 1900795), and the Open Research Fund of College Teacher Education, Zhejiang Normal University (Grant No. jykf20001).

[**] Corresponding Author: Peida Zhan, Emial: pdzhan@gmail.com

Individual growth and change has long been a focus of interest in educational, psychological, and behavioral studies. In recent decades, the learning diagnosis, which objectively quantifies learning status and provides diagnostic feedback, has drawn increasing interest. This approach aims to promote students' learning, based on the concept of "assessment for learning" (Wiliam, 2011) and on cognitive diagnostic assessment (Leighton & Gierl, 2007). Longitudinal learning diagnosis evaluates students' knowledge and skills and identifies their strengths and weaknesses over a period of time. The data collected from longitudinal learning diagnosis has provided researchers with opportunities to develop models for learning, which can not only be used to track individual growth over time but can also be used to evaluate the effectiveness of remedial teaching.

In recent years, several longitudinal learning diagnosis models (LDMs) or longitudinal cognitive diagnosis models have been proposed to provide theoretical support for this approach. Current longitudinal LDMs can be divided into two categories. The first are the latent transition analysis-based longitudinal LDMs (e.g., Chen, Culpepper, Wang, & Douglas, 2018; Collins & Wugalter, 1992; Kaya & Leita, 2017; Li, Cohen, Bottge, & Templin, 2016; Madison & Bradshaw, 2018; Wang, Yang, Culpepper, & Douglas, 2018; Zhang & Chang, 2019). The second group are the higher-order latent structural model-based longitudinal LDMs (e.g., de la Torre & Douglas, 2004; Huang, 2017; Lee, 2017; Zhan, Jiao, Liao, & Li, 2019a). The first group estimates the transition probabilities from one latent class or attribute to another or to the same latent class or attribute. The second group estimates the changes in higher-order latent ability over time, and from these it infers the changes in the lower-order latent attributes. Although the usefulness of these longitudinal LDMs in analyzing longitudinal learning diagnosis data has been evaluated through some simulation studies and a few applications, a small but very practical issue has not been

addressed. This is that the estimation strategy adopted by current longitudinal LDMs cannot provide timely diagnostic feedback, which is inconsistent with the idea of "assessment for learning" and therefore, with the fundamental purpose of formative assessment.

More specifically, the *simultaneity estimation strategy* is currently adopted by almost all longitudinal LDMs, and this involves the re-integration of response data from multiple time points into one large response matrix, which is then analyzed as a whole (Zhan et al., 2019a). The simultaneity estimation strategy is derived from traditional longitudinal psychological assessments that do not involve interventions. For example, researchers may be focusing on changes in children's intimacy with their parents over a period of time (Rice & Mulkeen, 1995), but here there are no interventions between the multiple time points at which the data are collected. This makes it impossible to guarantee any objective and accurate description of the changes as they occur. This strategy requires subjects to wait until all the tests are ended before an analysis of the results becomes available. Using this traditional longitudinal approach for learning assessment cannot, therefore, provide timely diagnostic feedback to either students or teachers. However, in many longitudinal learning diagnosis projects, both students and teachers would hope to know the degree of growth and the effectiveness of the remedial teaching as quickly as possible. In practice, highly motivated students may want to adjust their learning progress timeously on the basis of diagnostic feedback, while responsible teachers may want to adjust their teaching schedule in real-time based on the diagnostic feedback. From the perspective of test implementation, students may wish to exclude themselves from tests where they have already mastered the attributes assessed by those tests; while teachers can stop redundant remedial teaching when they discover that most of the students in a class have mastered the required attributes. Using the simultaneity parameter estimation strategy cannot meet

the need for timely diagnostic feedback that would facilitate these processes.

To obtain more timely diagnostic feedback, previous studies have adopted a *separated estimation strategy* to use the same cross-sectional model for an independent analysis of data at different time points (Wu, 2018). The inherent risk in using the separated estimation strategy is that the results of the diagnostic analysis at different time points may not be comparable, especially with the non-parallel tests that are commonly used in educational measurement. More specifically, a basic assumption in using the separated estimation strategy is that the invariance property (and especially, the item-invariance property[1]) holds in the cross-sectional model. If the invariance property holds, the estimates of the same sample of students at different time points will be comparable, and changes over time will reflect the actual changes in a student rather than artificial changes in the attribute metric (i.e., the meaning of mastery or nonmastery of an attribute; Bolt & Kim, 2018). Previous studies (Bradshaw & Madison, 2016; de la Torre & Lee, 2010; Ravand, Baghaei, & Doebler, 2020) have found that the invariance property may hold when the sample size is big enough and the model fits the data; however, perfect invariance is never observed.

As an alternative when using the separated estimation strategy, researchers can apply an anchor-item design, which ensures that anchor-items have consistent parameters across different time points (Xu & von Davier, 2008). This is described as the *anchor-item estimation strategy*. This strategy uses anchor-items to link different tests, and it thereby releases the assumption that the invariance property holds in the cross-sectional model.

Aiming to solve the practical problem caused by the simultaneity estimation strategy's inability to provide timely diagnostic feedback, this study proposes a new *Markov estimation strategy*, which assumes the Markov property (more details see

[1] Item-invariance property states that the examinee parameter (latent ability and latent attribute) estimates are independent of the particular set of items administered.

below). It should be emphasized that this study proposes an estimation strategy rather than a parameter estimation algorithm (such as the maximum likelihood estimation or the Bayesian MCMC estimation). The same parameter estimation algorithm can be applied to various estimation strategies, and the same estimation strategy can also be adopted with different parameter estimation algorithms.

For simplicity, but without loss of generality, a simple version of the longitudinal higher-order deterministic-inputs, noisy "and" gate (sLong-DINA) model (Zhan et al., 2019a) is used throughout this study. The rest of the paper starts with a review of the sLong-DINA model and of the simultaneity estimation strategy. In addition, the separated and anchor-item estimation strategies are briefly described. The proposed Markov estimation strategy is then presented, followed by a simulation study that evaluates and compares the psychometric properties of the four strategies described. Finally, the authors summarize their findings and discuss possible directions for future research.

Background

sLong-DINA Model

To make this study more focused, we take the sLong-DINA model as an example through which to illustrate the practical concerns described. The sLong-DINA model is a representative model of the higher-order latent structural model-based longitudinal LDMs. Differing from the complete version (i.e., the Long-DINA model), the special dimensions used to account for local item dependence among anchor items at different time points (Paek, Park, Cai, & Chi, 2014) are ignored in the sLong-DINA model to reduce model complexity and computational burden. The results from the empirical data analysis by Zhan et al. (2019a) also indicate that ignoring local item dependence in certain test scenarios may achieve better model-data fit.

If y_{nit} is the response of person n (n = 1, ..., N) to item i (i = 1, ..., I) at time point t (t = 1, ..., T), the sLong-DINA model can be expressed as follows:

first-order:

$$\text{logit}(P(y_{nit}=1\,|\,\boldsymbol{\alpha}_{nt},\gamma_{nm},\lambda_{0it},\lambda_{1it}))=\lambda_{0it}+\lambda_{1it}\prod_{k=1}^{K}\alpha_{nkt}^{q_{ikt}}, \quad (1)$$

second-order:

$$\text{logit}(P(\alpha_{nkt}=1\,|\,\theta_{nt},\xi_k,\beta_k))=\xi_k\theta_{nt}-\beta_k, \quad (2)$$

third-order:

$$\boldsymbol{\theta}_n=(\theta_{n1},\ldots,\theta_{nT})'\sim MVN_T(\boldsymbol{\mu},\boldsymbol{\Sigma}), \quad (3)$$

where $\boldsymbol{\alpha}_{nt}$ = (α_{n1t}, ..., α_{nKt})′ denotes person n's attribute profile at time point t, $\alpha_{nkt}\in\{0, 1\}$; λ_{0it} and λ_{1it} are the intercept and interaction parameter for item i at time point t, respectively; q_{ikt} is the element in an I-by-K Q-matrix at time point t; θ_{nt} is person n's general ability at time point t; ξ_k and β_k are the slope and difficulty parameters of attribute k on all time points, respectively, since the same latent structure is assumed to be measured at different time points; $\boldsymbol{\mu}$ = (μ_1, ..., μ_T)′ is the mean vector, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix.

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & & \\ \vdots & \ddots & \\ \sigma_{1T} & \cdots & \sigma_T^2\end{bmatrix},$$

where σ_{1T} is the covariance of the first and Tth general abilities. As a starting and reference point for subsequent time points, θ_{n1} is constrained to follow a standard normal distribution.

For a specific time point or T = 1, the sLong-DINA model is constrained to be the higher-order DINA (HO-DINA) model (de la Torre & Douglas, 2004). In this study, the HO-DINA model will be applied to the separated and anchor-item estimation strategies.

Note that there are two reasons why we did not consider using a general or

saturated model (Huang, 2017; Madison & Bradshaw, 2018). First, general models always need a large sample size to obtain a robust parameter estimate (Chiu, Sun, & Bian, 2018; Jiang & Ma, 2018; Ravand & Robitzsch, 2018). It is difficult to meet such a requirement in small-scale educational projects (such as school and classroom-level assessments). Second, the parameters in general models are often hard to interpret in practice (Ravand, 2016; Rojas, de la Torre, & Olea, 2012). Adequate parameter constraints are essential for obtaining interpretable and meaningful insights from the model, and these are especially important if educational and psychological applications are to meet the need for accountability.

Simultaneity Estimation Strategy

At present, no matter whether using the maximum likelihood estimation (Zhan et al., 2019a) or the full Bayesian MCMC estimation (Zhan, Jiao, Man, & Wang, 2019c), the sLong-DINA model adopts the simultaneity estimation strategy. The simultaneity estimation strategy means that, before the data analysis, $N \times I_t$ response matrices at T time points need to be merged into a $N \times \sum_{t=1}^{T} I_t$ longitudinal response matrix, and $I_t \times K$ Q_t-matrices at T time points also need to be merged into a $\sum_{t=1}^{T} I_t \times TK$ longitudinal Q-matrix (Zhan et al., 2019a). In such a case, the length of the estimated attribute pattern for each person is TK.

As shown in Figure 1, when a longitudinal learning diagnosis is assumed to contain four time points, the simultaneity estimation strategy is performed only after data collection at time point 4. Therefore, the simultaneity estimation strategy has at least two limitations. First, as mentioned, it cannot provide timely diagnostic feedback. Second, when the number of test points is large, the number of estimated parameters is excessive, and this may lead to a non-robust parameter estimation.

[Figure 1]

Separated and Anchor-Item Estimation Strategies

As shown in Figure 1, in comparison with the simultaneity estimation strategy, the separated estimation strategy is more straightforward and easy to understand, that is, the HO-DINA model or another cross-sectional model may be used independently for parameter estimation at different time points. While in essence, the separated estimation strategy and the anchor-item estimation strategy are almost the same, a significant difference is that the latter takes into account the requirements for anchor design and sets the item parameters of corresponding anchor-items at different time points so that they are equal, while freely estimating the other parameters (e.g., person parameters and latent structural parameters), as also shown in Figure 1.

The anchor-item estimation strategy is used specifically with anchor-item design tests, and new items appear at each follow-up time point. For repeated or parallel tests where every item is a potential anchor-item, the anchor-item estimation strategy is equivalent to the so-called fixed estimation strategy (Cho, Cohen, Kim, & Bottge, 2010; Wingersky & Lord, 1984) in which the item parameters from the first time point are calibrated and used in subsequent follow-up time points.

Markov Estimation Strategy

In general, the term Markov property refers to the memoryless property of a stochastic process. Specifically, a stochastic process has the Markov property if the conditional probability distribution of future states of the process depends only upon the present state and not on the sequence of events that preceded it.

Inspired by the Markov property, in the proposed Markov estimation strategy only the data from two adjacent time points are analyzed at any one time, and the second

time point of the current estimation will then be treated as the reference point for the next estimation. The estimated parameters of the second time point in the current estimation will be the fixed parameters of the next estimation. In other words, using the proposed strategy in longitudinal learning diagnosis indicates that the data analysis at time $t + 1$ is related only to the results at time t, but is unrelated to the results at time $t - 1$, or earlier.

As shown in Figure 1, the first data analysis can be performed after time point 1, and the parameter estimation results for time point 1 are obtained. Then, after time point 2, the second data analysis is carried out by simultaneity estimation including only time points 1 and 2. At this point, the parameter estimation results for time point 1 are fixed as "known values" to ensure that the parameter estimation results for these two time points are comparable. Then, after time point 3, a third data analysis is performed using a simultaneity estimation that includes only data from time points 2 and 3. At this point, the parameter estimation results for time point 2 are fixed as known values, and the results at time point 1 are not included. Finally, after time point 4, the simultaneity estimation for time points 3 and 4 are used for the fourth data analysis, and here the results of time point 3 are fixed as known values, and the results of time points 1 and 2 are no longer taken into consideration.

The estimation process can thus be explained as follows: (1) at time point t, all model parameters at time point $t - 1$ are fixed, including item parameters, latent structural parameters, and person parameters; (2) according to the construction logic of the sLong-DINA model, when estimating the general ability at time point t (θ_t), it is connected with θ_{t-1} as:

$$\theta_t = b_{t(t-1)}\theta_{t-1} + a_{t(t-1)} + \varepsilon_t, \tag{4}$$

where,

$$b_{t(t-1)} = \frac{\rho_{t(t-1)}\sigma_{\theta_t}}{\sigma_{\theta_{t-1}}}, \tag{5}$$

$$a_{t(t-1)} = \mu_{\theta_t} - b_{t(t-1)}\mu_{\theta_{t-1}}, \tag{6}$$

$$\varepsilon_t \sim N(0,\ \sigma_\varepsilon^2), \tag{7}$$

And where $\varrho_{t(t-1)}$ is the correlation coefficient between general abilities at two adjacent time points; σ_{θ_t} is the standard deviation of the general ability at time point *t*, which needs to be estimated; $\sigma_{\theta_{t-1}}$ is the standard deviation of the general ability at time point $t-1$, which is fixed at the value obtained from the estimation of time point $t-1$; μ_{θ_t} is the average of the general ability at time point *t*, which needs to be estimated; $\mu_{\theta_{t-1}}$ is the average of the general ability at time point $t-1$, which is fixed at the value obtained from the estimation of time point $t-1$; ε_t is the residual term at time point *t*, which follows the normal distribution with mean of 0 and variance of σ_ε^2, where $\sigma_\varepsilon^2 = \sigma_{\theta_t}(1-\rho_{t(t-1)}^2)$. Additionally, the prior distribution of general ability at the starting time point, which serves as the reference point, is constrained as $\theta_1 \sim N(0, 1)$.

In such cases, if there are *T* time points, the number of estimated parameters is $2K+2\sum_{k=1}^{K} I_t$, $2K+2\sum_{k=1}^{K}(I_t-m_t)+2M$, $2K+2\sum_{k=1}^{K}(I_t-m_t)+2M+3(T-1)$, and $2K+2\sum_{k=1}^{K}(I_t-m_t)+2M+{}^{T(T-1)}\!/_{2}+2(T-1)$ for the separated, anchor-item, Markov, and simultaneity estimation strategies, respectively, where *M* is the total number of anchor-items, and m_t is the number of anchor-items at each time point. However, the parameter estimation for the first three strategies is completed in *T* times, while the parameter estimation for the fourth strategy is completed only once. Thus, the simultaneity estimation strategy requires the maximum number of parameters to be estimated each time.

Compared with the simultaneity estimation strategy, the proposed Markov strategy ignores measurement error in the current estimation, which may affect the accuracy of

subsequent parameter estimations. However, as the proposed strategy needs to estimate only a fewer parameters each time, the robustness of its parameter estimation may be higher than that of the simultaneity estimation strategy. More importantly, using the Markov strategy allows students and teachers to utilize diagnostic feedback more timeously.

Furthermore, in comparison with the separated estimation strategy, the Markov strategy considers the connection between different time points, that is, it estimates the model parameters of the current time point based on the fixed model parameters of the previous time point. Theoretically, the proposed strategy thus makes the model parameters at different time points comparable. In addition, although the anchor-item estimation strategy does consider the connection between different time points, it only constrains the anchor-items at the different time points to have time invariance. Since there are no assumptions that latent structural parameters (such as attribute slope and attribute difficulty) will be cross-time invariant, this strategy may result in general abilities not being comparable across time points.

As mentioned before, the purpose of this study is to propose a parameter estimation strategy that meets the need for timely feedback in longitudinal learning diagnosis projects. Based on the introduction given to the four strategies, the basic assumptions of this study are as follows: (1) the Markov estimation strategy and the simultaneity estimation strategy have similar diagnostic results, that is, the diagnostic consistency of them is high; (2) the Markov estimation strategy can provide more accurate diagnostic results than the separated or anchor-item estimation strategies; (3) the performance of the anchor-item estimation strategy is better than that of the separated estimation strategy; (4) the robustness of the parameter estimation in the Markov estimation strategy is higher than that in the simultaneity estimation strategy.

A simulation study was conducted to explore the differences in the diagnostic

results of the four strategies under simulated conditions.

Simulation Study

Design and Data Generation

In the simulation study, three factors were manipulated. First, two levels of sample sizes were considered, N = 200 and 400. According to the national situation in the authors' country, sample sizes of 200 and 400 translate to approximately 5 and 10 classes with 40 students in each. In school-level projects, more classes and more students per class are rare. Then, three levels for the number of time points were considered, T = 2, 3, and 4. Third, two levels for the number of items at each time point were considered, I_t = 15 and 30.

Additionally, four attributes (K = 4) were measured. The first two items for I_t =15 and the first four items for I_t = 30 were used as anchor-items. The simulated Q-matrices are presented in Figure 2. In practice, it is common to used high-quality items as anchor items, thus the anchor item parameters were fixed as λ_{0it} = –2.197 and λ_{1it} = 4.394. In such a case, aberrant response probabilities (i.e., guessing and slipping) are approximately equal to 0.1. In addition, non-anchor item parameters were generated from a bivariate normal distribution with a negative correlation coefficient as follows:

$$\begin{pmatrix} \lambda_{0it} \\ \lambda_{1it} \end{pmatrix} \sim MVN_2\left(\begin{pmatrix} -2.197 \\ 4.394 \end{pmatrix}, \begin{pmatrix} 1.0 & -0.6 \\ -0.6 & 1.0 \end{pmatrix}\right). \quad (8)$$

This setting leads the guessing and slipping probabilities for all items to follow a positively skewed distribution (mean ≈ 0.1, minimum ≈ 0.01, and maximum ≈ 0.6). Assuming that guessing and slipping parameters follow a negative correlation is more realistic (Zhan, Jiao, Liao, & Bian, 2019b). Attribute difficulty parameters were fixed as $\boldsymbol{\beta}$ = (–1, –0.5, 0.5, 1)′. The correlation among the general abilities at different time points was set as 0.9. Between two consecutive time points, the overall mean growth was set at

0.5, and the overall scale change was set as $\sqrt{1.25}$. The general abilities at T time points were generated from a T-way multivariate normal distribution according to Equation 3. At each time point, the true attribute pattern for each person was generated according to Equation 2. Finally, the response data were generated from y_{nit} ~ Bernoulli ($P(y_{nit}=1)$), where $P(y_{nit}=1)$ was defined as in Equation 1.

Analysis

In this study, the parameters of the sLong-DINA model were estimated using the full Bayesian MCMC estimation and using the JAGS (Just Another Gibbs Sampler) software. The corresponding parameter estimation code is also available from the authors. More details about how to use the JAGS for Bayesian CDM estimation can be found in a tutorial by Zhan et al. (2019c).

For the separated estimation strategy, the HO-DINA model was used independently for parameter estimation at each time point. For the anchor-item estimation strategy, in addition to using the HO-DINA model for parameter estimation at each time point, the same anchor-items at different time points were assumed to have consistent item parameters. For the simultaneity estimation strategy, response data at multiple time points were re-integrated into a longitudinal response matrix, and then the matrix was jointly analyzed at once by using the sLong-DINA model. For the Markov estimation strategy, the analysis process followed Equations 4 to 7. Additionally, it should be noted that in all four strategies, the prior distribution of general ability at the starting time point, which serves as the reference point, was constrained as $\theta_1 \sim N(0, 1)$. The prior distribution of the remaining model parameters are shown in the Appendix.

Fifty replications were implemented for each condition. For each replication, two Markov chains with random starting points were used. For the simultaneity estimation

strategy, in each chain, 25,000 iterations were run, with the first 20,000 iterations discarded as burn-in. In contrast, for the other three strategies, 15,000 iterations were run in each chain, with the first 10,000 iterations discarded as burn-in. The remaining 10,000 iterations (5,000 in each chain) were utilized for the model parameter inferences. The potential scale reduction factor (PSRF; Brooks & Gelman, 1998) was computed to assess the convergence of all parameters. PSRF values of less than 1.1 or 1.2 indicate convergence. Our study indicated that PSRF was generally less than 1.01, suggesting acceptable convergence for the setting specified.

To evaluate parameter recovery, the bias and the root mean square error (RMSE) were computed as $\text{bias}(\hat{v}) = \sum_{r=1}^{R} \frac{\hat{v}_r - v}{R}$ and $\text{RMSE}(\hat{v}) = \sqrt{\sum_{r=1}^{R} \frac{(\hat{v}_r - v)^2}{R}}$, where $\hat{v}$ and v are the estimated and true values of the model parameters, respectively; and R is the total number of replications. In addition, the correlation between the true values and estimated values (Cor) for the parameters were computed to evaluate the recovery. For attribute recovery, the attribute and pattern correct classification rate (i.e., ACCR and PCCR) were computed to evaluate the classification accuracy of individual attributes and profiles: $\text{ACCR} = \sum_{r=1}^{R} \sum_{n=1}^{N} \text{I}(\hat{\boldsymbol{\alpha}}_{nkr} = \boldsymbol{\alpha}_{nkr}) / NR$ and $\text{PCCR} = \sum_{r=1}^{R} \sum_{n=1}^{N} \text{I}(\hat{\boldsymbol{\alpha}}_{nr} = \boldsymbol{\alpha}_{nr}) / NR$, where I(·) is an indicator function.

If two strategies have the same correct classification rate (unless it is 1), it does not mean that their classification results are consistent. Therefore, we also used a classification consistency index (CCI) to evaluate the degree of classification consistency of the four strategies:

$$\text{CCI} = \frac{\sum_{r=1}^{R} \sum_{n=1}^{N} \text{I}(\hat{\boldsymbol{\alpha}}_{nr} = \hat{\boldsymbol{\alpha}}_{nr}^{*})}{NR}, \tag{9}$$

where $\hat{\boldsymbol{\alpha}}_{nr}^{*}$ is the estimated attribute pattern of person n in replication r when the benchmark method is adopted; $\hat{\boldsymbol{\alpha}}_{nr}$ is the estimated attribute pattern of person n in replication r when another method is adopted. CCI = 1 indicates that the diagnostic

results of the two methods are completely consistent, and CCI = 0 indicates complete inconsistency.

Results

Figure 3 displays the recovery of item parameters in the four strategies. First, the test length has little effect on the recovery of item parameters. Second, the four strategies show a consistent pattern across different conditions, for example, with an increased sample size, the recovery of item parameters becomes better, while the number of time points has limited effect on the recovery of item parameters. The main focus of this study was on the performance differences of the four strategies, and it was found that: (1) their performances were similar; (2) the performance of the simultaneity estimation strategy was the best, the Markov estimation strategy was the second best, followed by the anchor-item estimation strategy and the separated estimation strategy with a very small difference between the last two.

Figure 4 presents the posterior standard deviation (i.e., the standard error) of the item parameters produced by the four strategies to reflect the robustness of their item parameter estimation. The variation tendency of the intercept and interaction parameters across different manipulated factors is basically the same. At the overall level, for both intercept and interaction parameters, the posterior standard deviation produced by the Markov estimation strategy is the smallest, followed by the anchor-item estimation strategy, and the simultaneity estimation strategy. The separated estimation strategy produced the largest posterior standard deviation. When $t \geq 2$, the variation tendency of each parameter at each time point was consistent with that at the overall level. Only at the first time point, i.e., $t = 1$, the posterior standard deviation produced by the simultaneity estimation strategy was smaller than the others. In brief, it was found that in item parameter estimation the overall robustness of the

Markov estimation strategy is a little higher than that of the simultaneity estimation strategy.

Figure 5 presents the recovery of general abilities in the four strategies. First, the test length has little effect on the recovery of general abilities. Second, the four strategies show a consistent pattern under different simulation conditions. For any strategy, the general ability at a particular time point is almost unaffected by changes in simulation conditions. In addition, it can be seen that in all conditions, (1) the performance of the four strategies is quite different; (2) the performance of the simultaneity estimation strategy is the best, followed by the Markov estimation strategy, and followed by the anchor-item estimation strategy and the separated estimation strategy, the last two showing little difference; (3) according to RMSE and Cor, the comparison between the simultaneity estimation strategy and the Markov estimation strategy shows that the advantages of the former are mainly reflected in the early stages of the longitudinal test (e.g. at time point 1); (4) with additional time points, the recoveries for the anchor-item and separated estimation strategies become worse. According to bias, they tend to underestimate the general abilities in the later stages of the longitudinal test.

Figure 6 displays the posterior standard deviation in the general abilities as produced by the four strategies to reflect the robustness of their general ability parameter estimation. First, since there is a lack of connection between general abilities at various time points, the posterior standard deviation of general abilities produced by the anchor-item and separated estimation strategies is much larger than those produced by the simultaneity and Markov estimation strategies. Second, at the overall level, the posterior standard deviations of general abilities produced by the simultaneity and Markov estimation strategies are basically the same, with the former slightly less than the latter. When compared with the other three strategies, the main

advantage of the simultaneity estimation strategy is reflected at time point 1.

Figure 7 displays the recovery of attributes in the four strategies, which is often the most important factor in learning diagnosis. First, test length has a great influence on the recovery of attributes, especially on the PCCR. With increasing test length, ACCR and PCCR both improve. Second, the ACCR of the four strategies is higher across different simulation conditions. It should be noted that the PCCR in Figure 7 focuses on whether all *TK* attributes can be correctly recovered (e.g., if $T = 4$, the pattern contains 16 attributes), which is known as the longitudinal PCCR (Zhan et al., 2019a). It is shown that (1) with increased time points, the PCCR of the four strategies declined slightly, but it remained above 0.8 for $I = 30$ conditions and above 0.6 for $I = 15$ conditions; (2) the PCCR is highest for the simultaneity estimation strategy, followed by the Markov estimation strategy (about 1% lower than the former), and the anchor-item estimation strategy and separated estimation strategy with little difference between them.

The simulation results show that the simultaneity estimation strategy is relatively optimal. Therefore, in this study, the classification results for the simultaneity estimation strategy were taken as the benchmark for the CCI. Figure 8 shows the consistency of the diagnostic results for the four strategies First, the consistency of the four strategies was high in both the short test and the long test, and the CCI was further improved with the increase in test length. Second, it can be seen that (1) with the increase in time points, the CCI in all four strategies decreased slightly; (2) the CCI of the Markov estimation strategy is relatively the highest, followed by the anchor-item and the separated estimation strategies with a small difference. In other words, the results indicate that the Markov and the simultaneity estimation strategies have high diagnostic consistency.

Overall, the results of the simulation study show that: (1) the model parameter recovery of the four strategies is good. In particular, in ACCR and PCCR, the most

important indices in learning diagnosis, the difference between the four strategies is only about 1%; (2) the diagnostic results of the four strategies are highly consistent; (3) the performances of the four strategies are presented in the following order: simultaneity > Markov > anchor-item ≥ separated; (4) the robustness of parameter estimation in the anchor-item and separated estimation strategies are worse than those in the simultaneity and Markov estimation strategies; and in comparison with the simultaneity estimation strategy, the overall robustness of the Markov estimation strategy is slightly higher for item parameters but slightly lower for general abilities.

[Figures 2 to 8]

Conclusion and Discussion

To meet the need for timely diagnostic feedback in practical longitudinal learning diagnosis, this study proposed a new parameter estimation strategy, the Markov estimation strategy. Compared with the simultaneity estimation strategy, the major practical advantage of the proposed strategy is that it allows students and teachers to adjust their follow-up learning and teaching plans timeously, according to a current diagnosis. A simulation study was conducted to explore and compare the performance of four estimation strategies, namely, the simultaneity, the Markov, the anchor-item, and the separated estimation strategies. The results show that their performance is highly consistent, and the following relative order is presented: simultaneity > Markov > anchor-item ≥ separated. Overall, although accuracy in parameter estimation is sacrificed slightly with the proposed strategy, it has the advantage of providing timely diagnostic feedback, which is in line with the idea of "assessment for learning" and the needs of formative assessment.

The study is therefore able to make the following practical suggestions:

(1) If the purpose of longitudinal learning diagnosis is to pursue accuracy in the diagnostic results (e.g., for high-stakes tests), the simultaneity estimation strategy is recommended.

(2) If the purpose of the longitudinal learning diagnosis is to promote the students' learning or to adjust current teaching (i.e., for low-stakes tests), the Markov estimation strategy is recommended.

(3) If the purpose of the longitudinal learning diagnosis is to quickly and conveniently grasp students' learning status and development trends (i.e., there is no special requirement for diagnostic accuracy), the anchor-item or the separated estimation strategies can be adopted.

In our view, in practice, it is more valuable to exchange the accuracy of a model parameter estimation for timeliness in test feedback. This study provides theoretical support for promoting the application of "assessment for learning" in psychological and educational assessments, and it also provides methodological support for formative assessments.

Despite these promising results, further studies are needed. First, this study explores the performance of only one longitudinal LDM (i.e. the sLong-DINA model) under the four strategies. Whether the conclusions would apply equally to other longitudinal LDMs warrants further study. Second, the number of simulation conditions in this study was limited. More independent variables (e.g., the number of attributes and the attribute hierarchies) and more complex test situations (e.g., different types of missing data) could be considered in future studies to provide more reference information for practitioners. Third, based on the proposed Markov estimation strategy, further studies could attempt to incorporate the intervention information (e.g., different remedial teaching methods and the number of attributes a student has mastered currently) after each item point in the follow-up estimations (Park, Xing, & Lee, 2018;

Wang et al., 2018). Fourth, in Bayesian estimation, the prior distribution reflects the beliefs of the data analyst, and in this study, the prior distributions selected are shown in the Appendix. The posterior distribution of model parameters will be affected by their prior distribution, especially for a small sample size or a limited number of items. In practice, we recommend that the data analyst select appropriate prior distributions based on the actual situation rather than copy those given in the appendix.

References

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing, 16*, 99–118.

Bolt, D. M., & Kim, J.-S. (2018). Parameter invariance and skill attribute continuity in the DINA model. *Journal of Educational Measurement, 55*, 264–280.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2017). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement, 42*, 5–23.

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika, 83*, 355–375.

Cho, S.-J., Cohen, A. S., Kim, S.-S., & Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Applied Psychological Measurement, 34*, 384–504.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*, 131–157.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement, 47*, 115–127.

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement, 54*, 440–480.

Jiang, Z., & Ma, W. (2018). Integrating differential evolution optimization to cognitive diagnostic model estimation. *Frontiers in Psychology, 9*:2142.

Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement, 77*, 369–388.

Lee, S. Y. (2017). *Growth curve cognitive diagnosis models for longitudinal assessment*. Unpublished doctoral dissertation, University of California, Berkeley.

Leighton, J. P., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Li, F., Cohen, A., Bottge, B, & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement, 76*(2), 181–204

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika, 83*, 963–990.

Paek, I., Park, H.-J., Cai, L., & Chi, E. (2014). A comparison of three IRT approaches to examine ability change modeling in a single-group anchor test design. *Educational and Psychological Measurement, 74*, 659–676.

Park, Y. S., Xing, K., & Lee, Y.-S. (2018). Explanatory cognitive diagnostic models: Incorporating latent and observed predictors. *Applied Psychological Measurement*, 42(5), 376–392.

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choices: A study of reading comprehension. *Educational Psychology*, 38, 1255–1277.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*, 782–799.

Ravand, H., Baghaei, P., & Doebler, P. (2020). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology, 10*:2930.

Rice, K. G., & Mulkeen, P. (1995). Relationships with parents and peers: A longitudinal study of adolescent intimacy. *Journal of Adolescent Research, 10*(3), 338-357.

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics, 43*, 57–87

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37, 3–14.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347–364

Wu, H.-M. (2018). Online individualised tutor for improving mathematics learning: A cognitive diagnostic model approach. *Educational Psychology*. Advanced Online Publication. Retrieved from https://doi.org/10.1080/01443410.2018.1494819

Xu, X., & von Davier, M. (2008). *Linking for the general diagnostic model*. ETS Research Report Series (ETS RR-08-08), ETS.

Zhan, P., Jiao, H., Liao, D., & Li, F. (2019a). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics, 44*, 251–281.

Zhan, P., Jiao, H., Liao, M., & Bian, Y. (2019b). Bayesian DINA modeling incorporating within-item characteristics dependency. *Applied Psychological Measurement, 43*, 143–158.

Zhan, P., Jiao, H., Man, K, & Wang, L. (2019c). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics, 44*, 473–503.

Zhang, S., & Chang, H. (2019). A multilevel logistic hidden Markov model for learning

under cognitive diagnosis. *Behavior Research Methods*. Advanced Online Publication. Retrieved from https://doi.org/10.3758/s13428-019-01238-w

Appendix

The prior distributions used in the current study:

1. Simultaneity estimation strategy

$$\boldsymbol{\theta}_n \sim \mathrm{MVN}_T(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mu_{t\geq 2} \sim N(0,1), \boldsymbol{\Sigma} = \boldsymbol{\Delta}\boldsymbol{\Delta}' \text{ (Cholesky transformation)},$$

$$\boldsymbol{\Delta} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \varphi & \psi & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \varphi & \varphi & \cdots & \psi \end{pmatrix}, \varphi \sim \mathrm{Norm}(0,1), \psi \sim \mathrm{Gamma}(1,1),$$

$$\xi_k \sim \mathrm{Norm}(0,4), \beta_k \sim \mathrm{Norm}(0,4)\ \ \mathrm{I}(\beta_k > 0),$$

$$\lambda_{0it} \sim \mathrm{Norm}(-2.197, 4), \lambda_{1it} \sim \mathrm{Norm}(4.394, 4)\ \ \mathrm{I}(\lambda_{1it} > 0).$$

2. Markov estimation strategy

As described in the main text, the main difference between the Markov estimation strategy and the simultaneity estimation strategy lies in the estimation of general ability at time point t based on the fixed value of general ability at time point $t - 1$ (see Equations 5 to 8). The prior distribution of each parameter is:

$$\rho_{t(t-1)} \sim \mathrm{Uniform}(0.5,1), \sigma_{\theta_t} \sim \mathrm{Norm}(1,1), \mu_{\theta_t} \sim \mathrm{Norm}(0,1).$$

The prior distribution of the other parameters is the same as above.

3. Anchor-item and separated estimation strategies

There is no difference between the two estimation strategies in setting the prior distribution of parameters. They differ from the Markov estimation strategy in that the anchor-item estimation strategy does not consider the connection between general abilities at adjacent time points, and there are no fixed latent structural parameters. Therefore, the prior distribution is determined as:

$$\theta_{t\geq 2} \sim \mathrm{Norm}(\mu_\theta, \sigma_\theta^2), \mu_\theta \sim \mathrm{Norm}(0,1), \sigma_\theta^2 \sim \mathrm{InvGamma}(1,1),$$

$$\xi_{kt} \sim N(0,4), \beta_{kt} \sim N(0,4)\ \ \mathrm{I}(\beta_k > 0).$$

The prior distribution of the other parameters is the same as above.

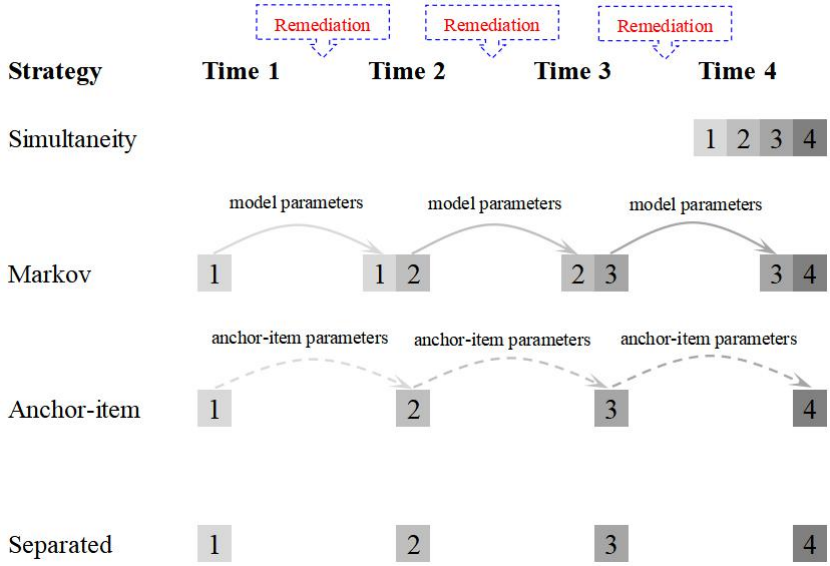
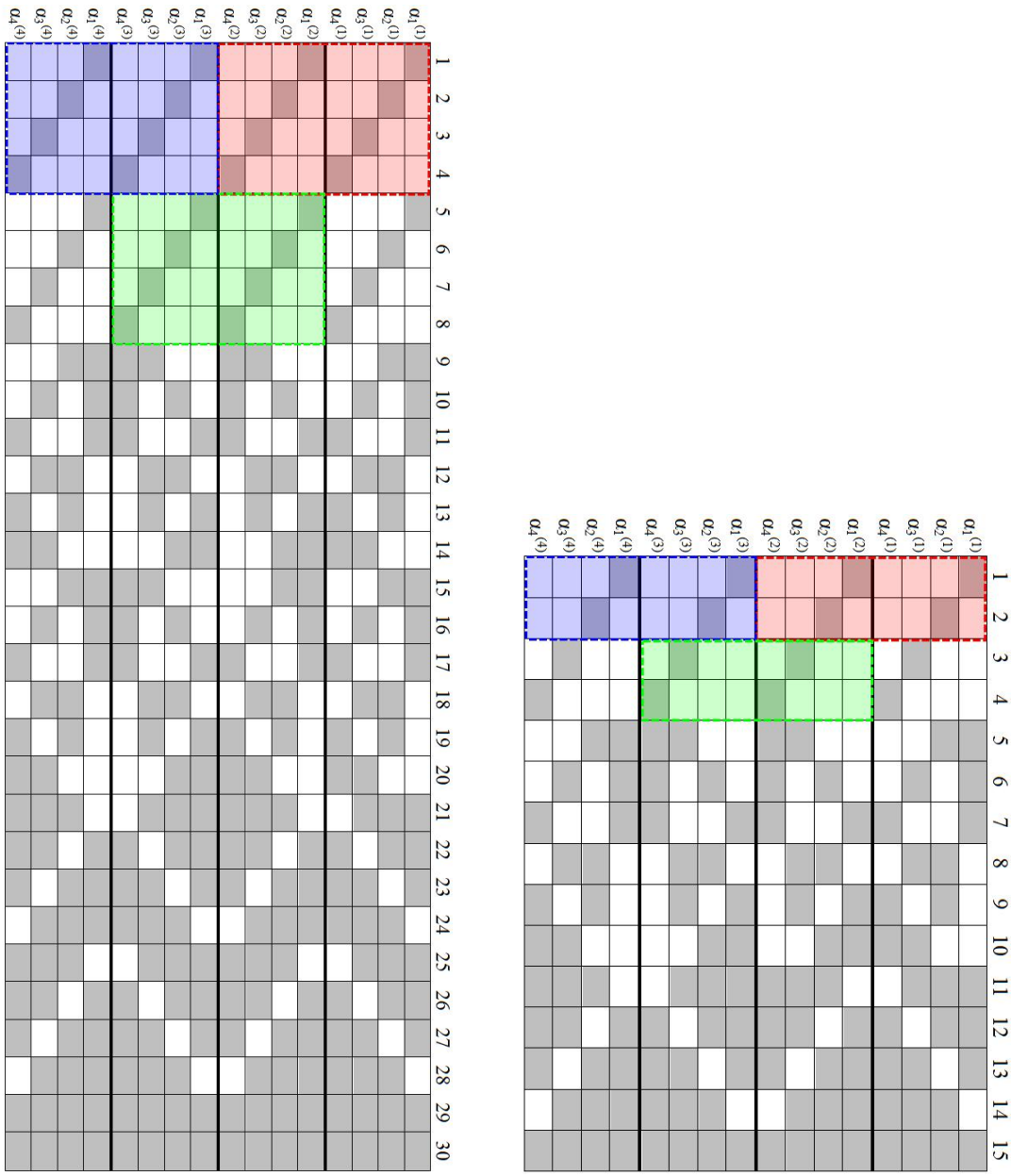


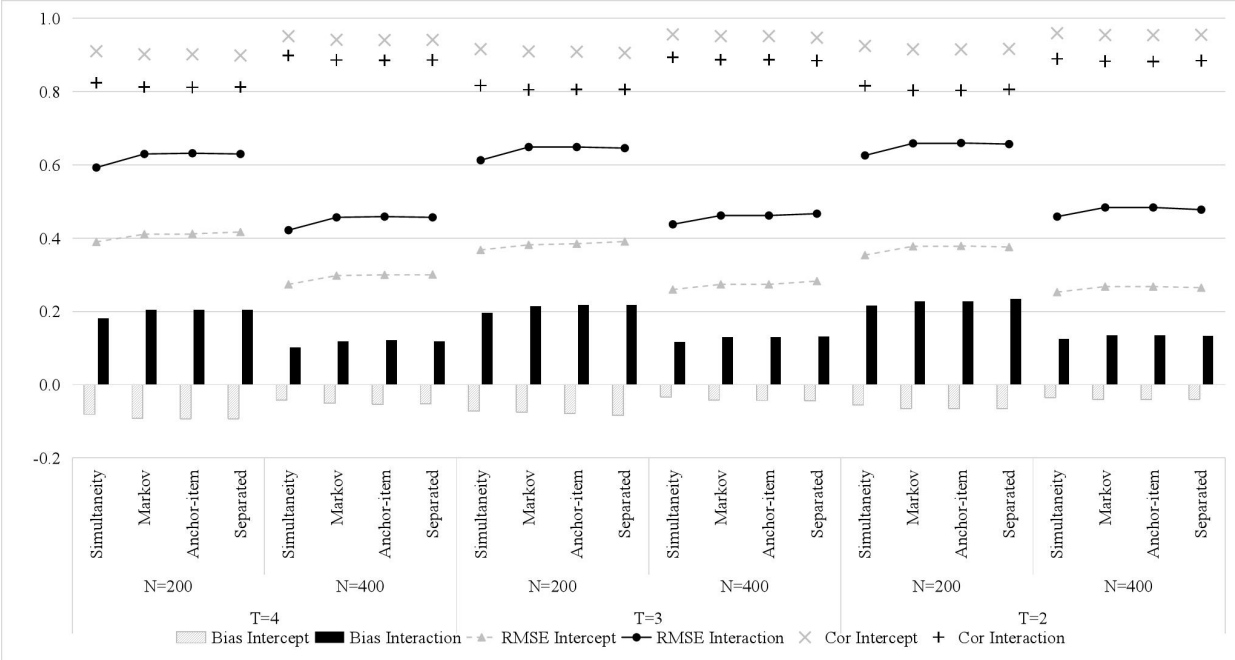
Figure 1. Four estimation strategies for the sLong-DINA model



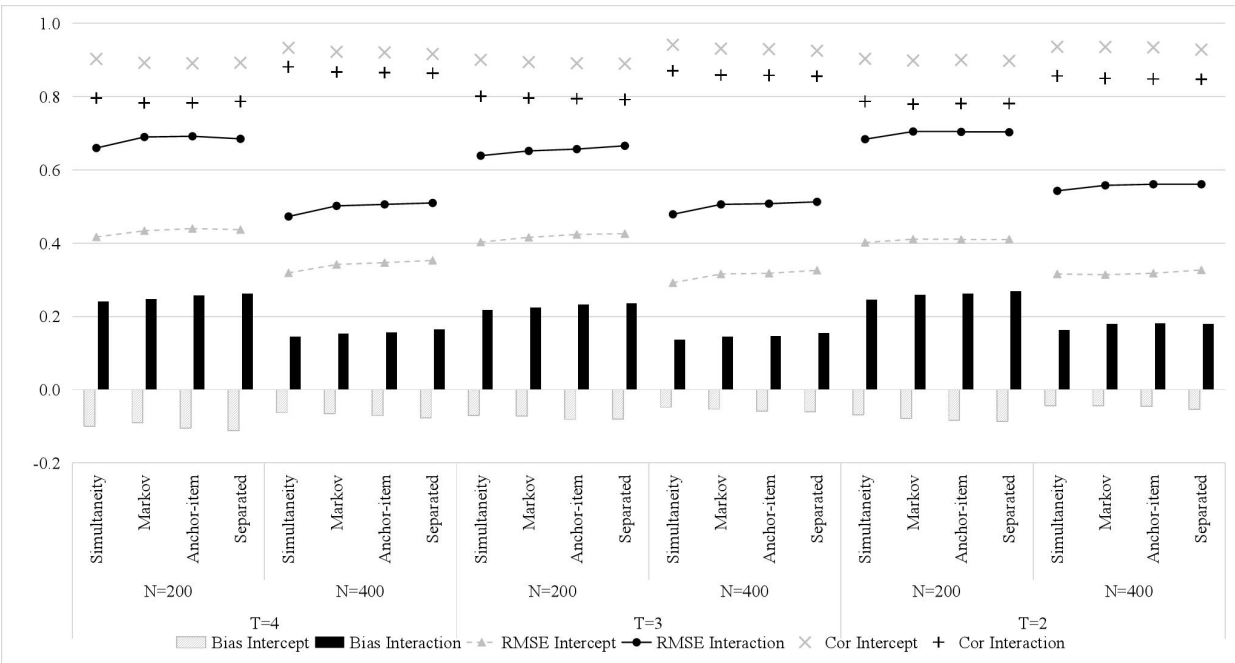
(a) I =30 **(b) I = 15**

Figure 2. Q-matrices in the simulation study

Notes: the same color blocks indicate that tests have the same anchor-items; I = test length.



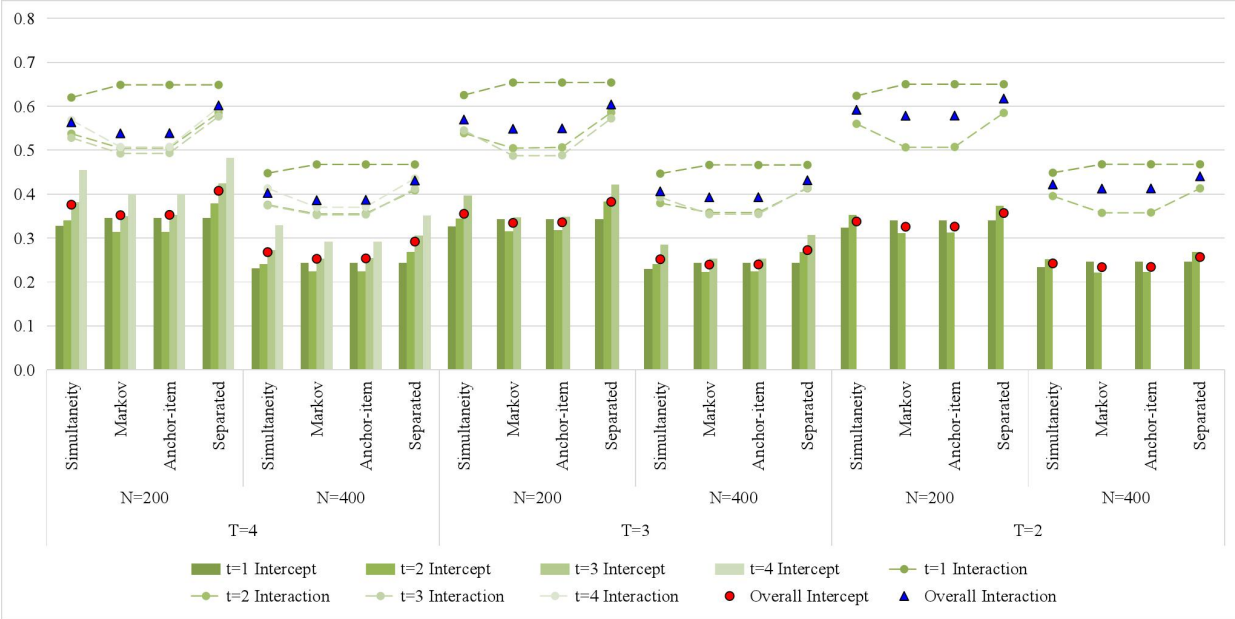
(a) I = 30



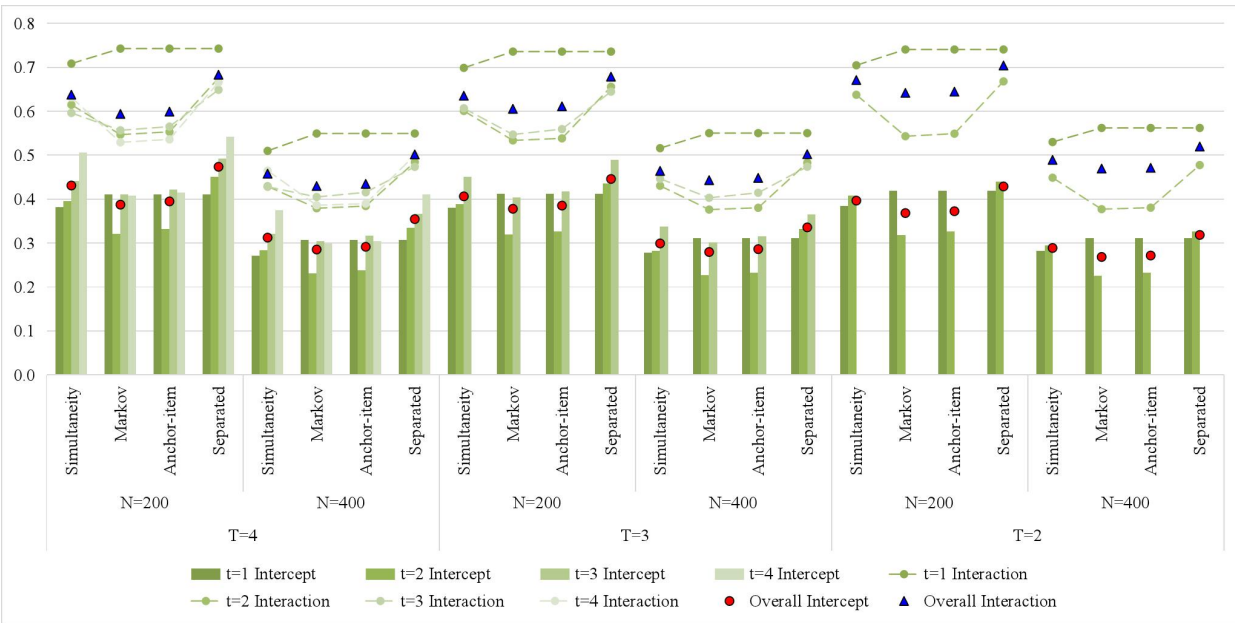
(b) I = 15

Figure 3. The recovery of item parameters in the four estimation strategies

Notes: T = time point; N = sample size; I = test length; RMSE = root mean square error; Cor = correlation between true and generated values; Intercept = item intercept; Interaction = item interaction; Simultaneity = simultaneity estimation strategy; Markov = Markov estimation strategy; Anchor-item = anchor-item estimation strategy; Separated = separated estimation strategy.



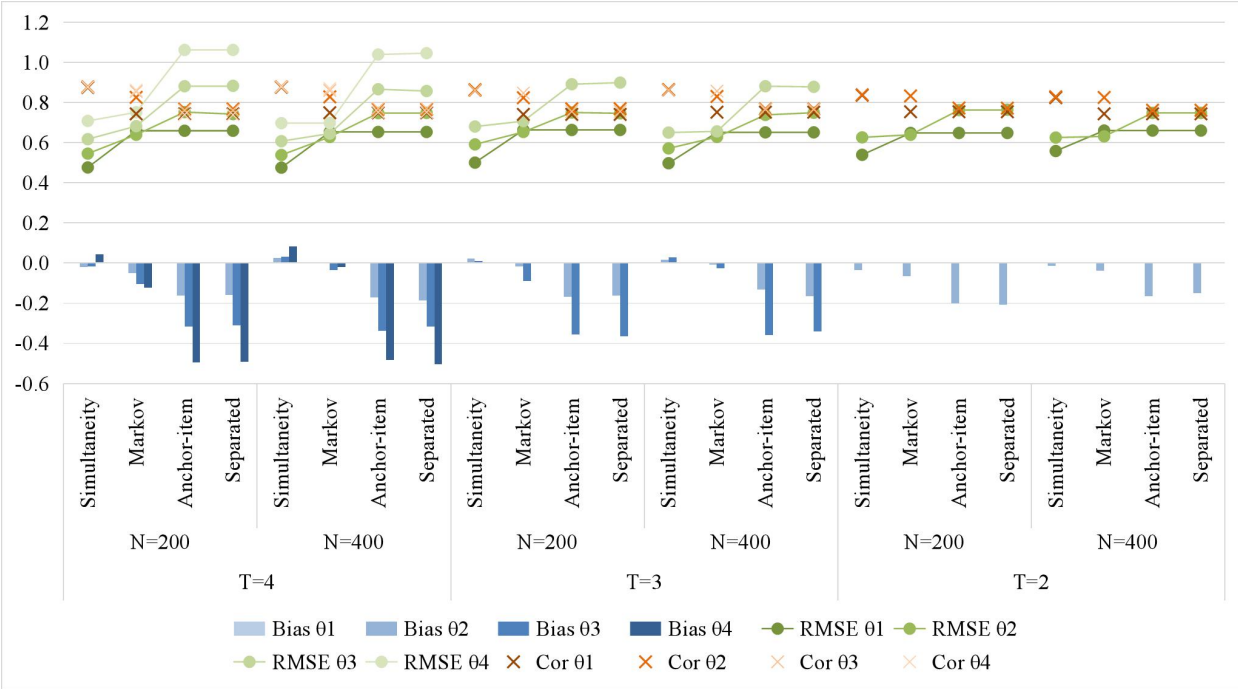
(a) I = 30



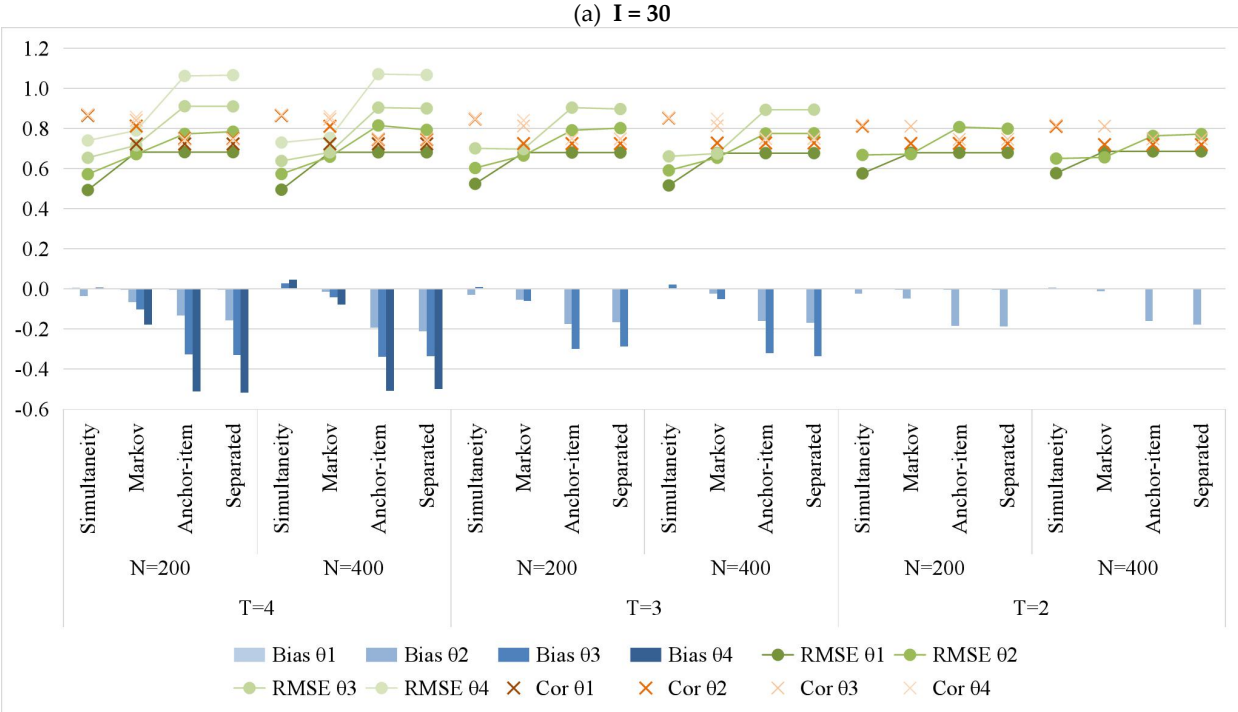
(b) I =15

Figure 4. The posterior standard deviation of item parameters in the four estimation strategies

Notes: T = time point; N = sample size; I = test length; t = *t*-th time point; Overall = average of the values at T time points; Intercept = item intercept; Interaction = item interaction; Simultaneity = simultaneity estimation strategy; Markov = Markov estimation strategy; Anchor-item = anchor-item estimation strategy; Separated = separated estimation strategy.



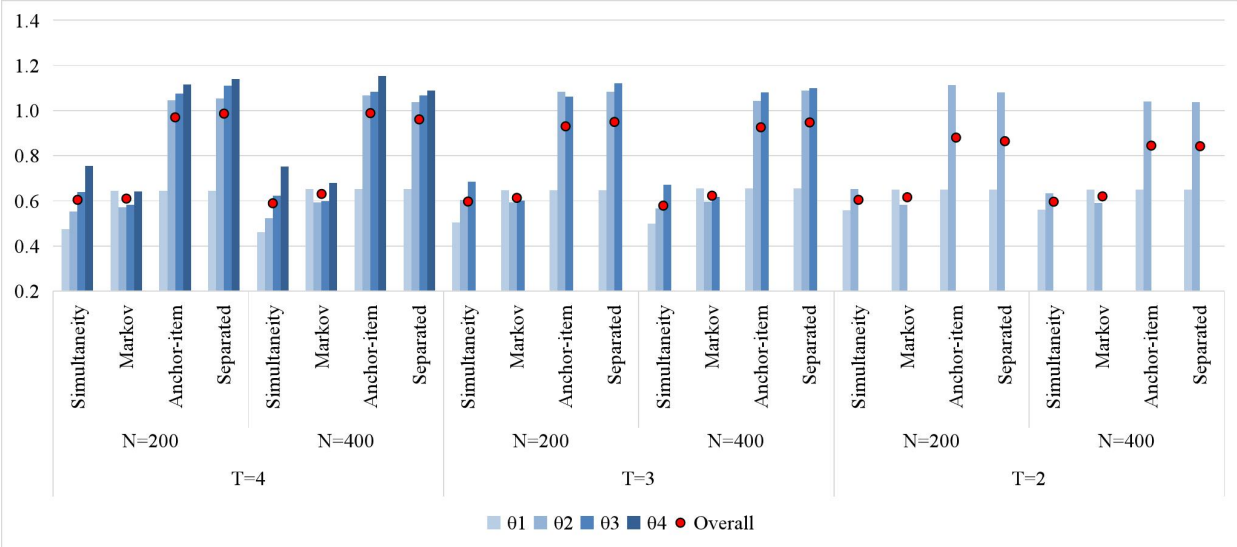
(a) **I = 30**



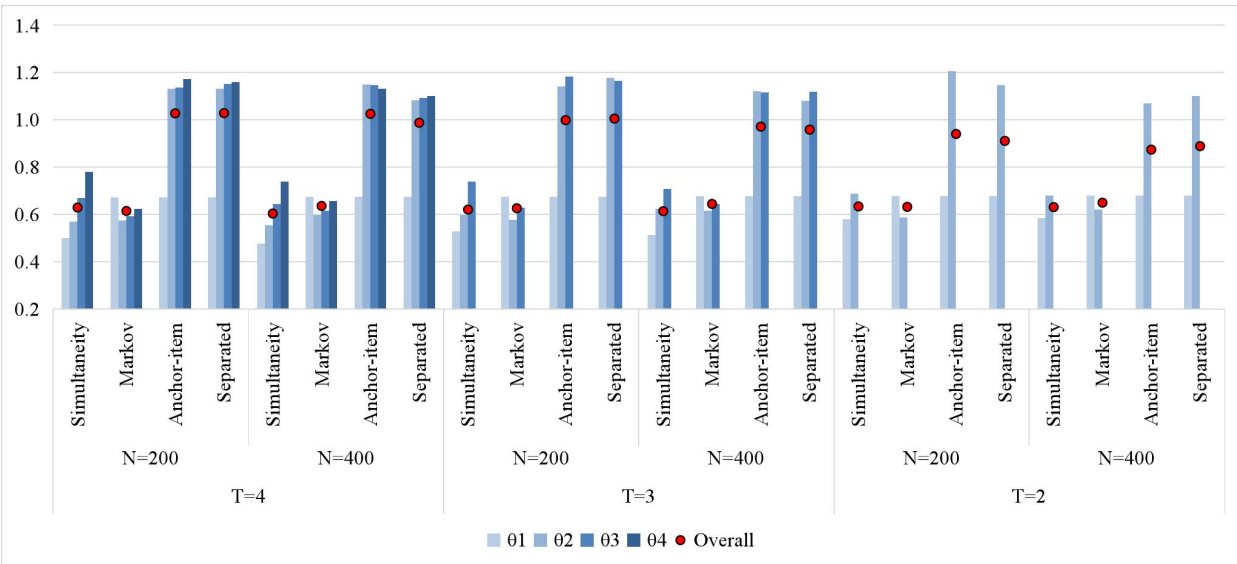
(c) I =15

Figure 5. The recovery of general abilities in the four estimation strategies

Notes: T = time point; N = sample size; I = test length; RMSE = root mean square error; Cor = correlation between true and generated values; θ = general ability; Simultaneity = simultaneity estimation strategy; Markov = Markov estimation strategy; Anchor-item = anchor-item estimation strategy; Separated = separated estimation strategy.



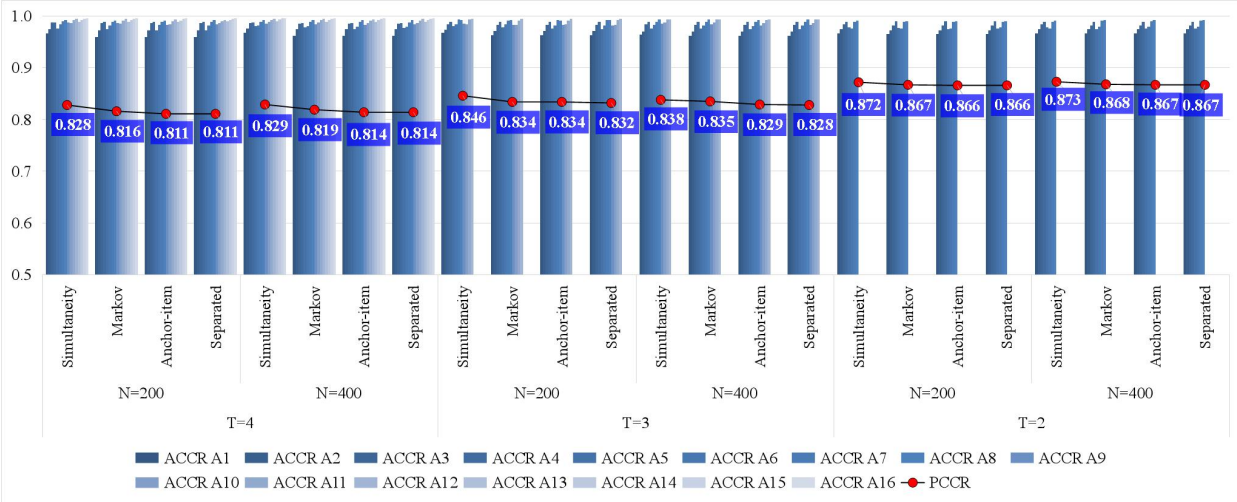
(a) I = 30



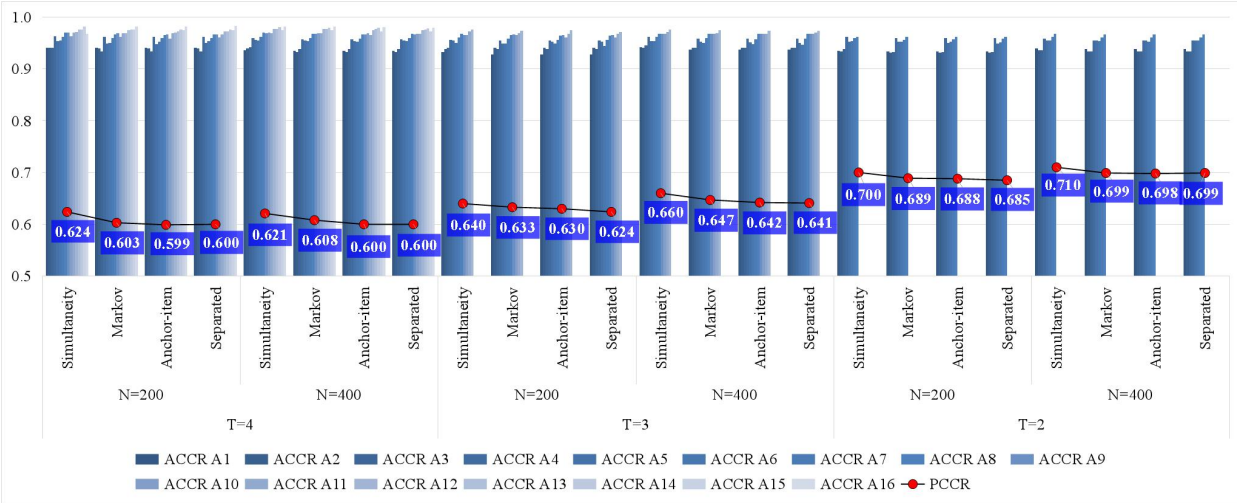
(b) I = 15

Figure 6. The posterior standard deviation of general abilities in the four estimation strategies

Notes: T = time point; N = sample size; I = test length; θ = general ability; Overall = average of the values at T time points; Simultaneity = simultaneity estimation strategy; Markov = Markov estimation strategy; Anchor-item = anchor-item estimation strategy; Separated = separated estimation strategy.



(a) I = 30



(b) I = 15

Figure 7. The recovery of attributes in the four estimation strategies

Notes: T = time point; N = sample size; I = test length; RMSE = root mean square error; Cor = correlation between true and generated values; ACCR = attribute correct classification rate; PCCR = pattern correct classification rate; Simultaneity = simultaneity estimation strategy; Markov = Markov estimation strategy; Anchor-item = anchor-item estimation strategy; Separated = separated estimation strategy.

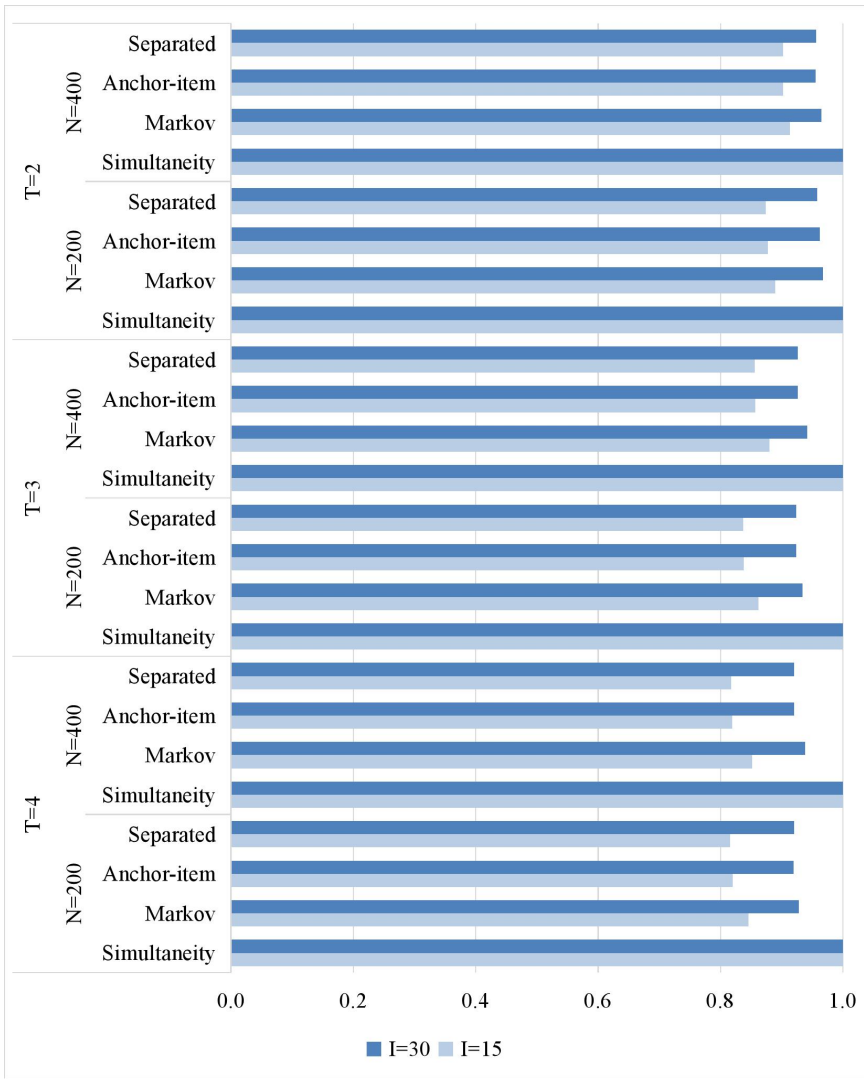


Figure 8. The classification consistency index among the four estimation strategies

Notes: the results of the simultaneity estimation strategy were used as the baseline; T = time point; N = sample size; I = test length; Simultaneity = simultaneity estimation strategy; Markov = Markov estimation strategy; Anchor-item = anchor-item estimation strategy; Separated = separated estimation strategy.